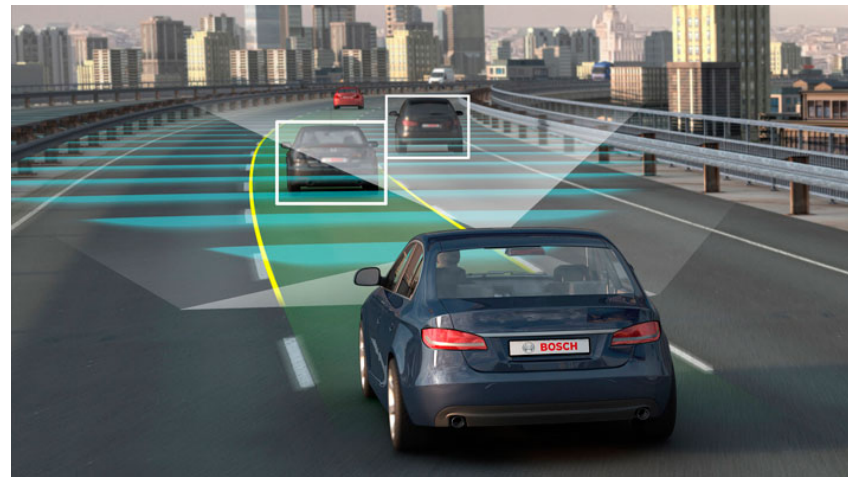
Recognition of Activity

Cees G.M Snoek & Arnold W.M Smeulders
University of Amsterdam

MOTIVATION

The many faces of video.

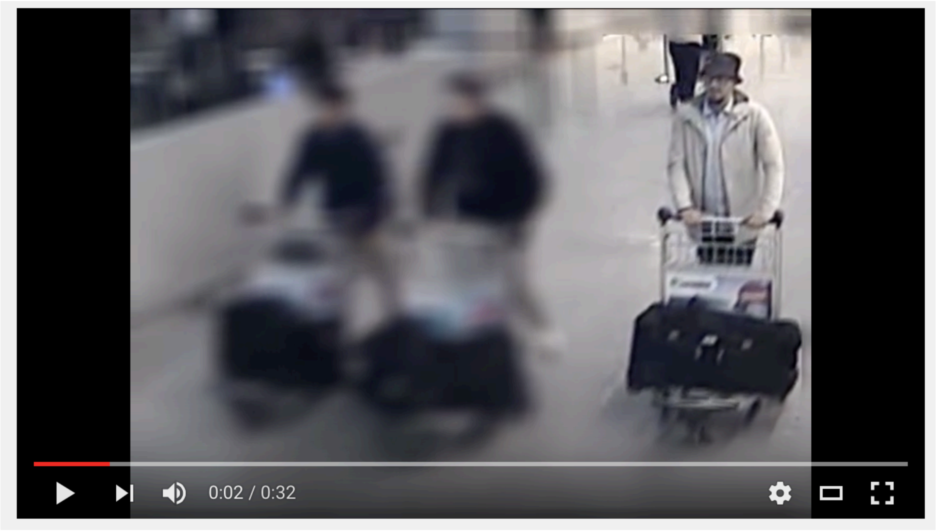
New purposes



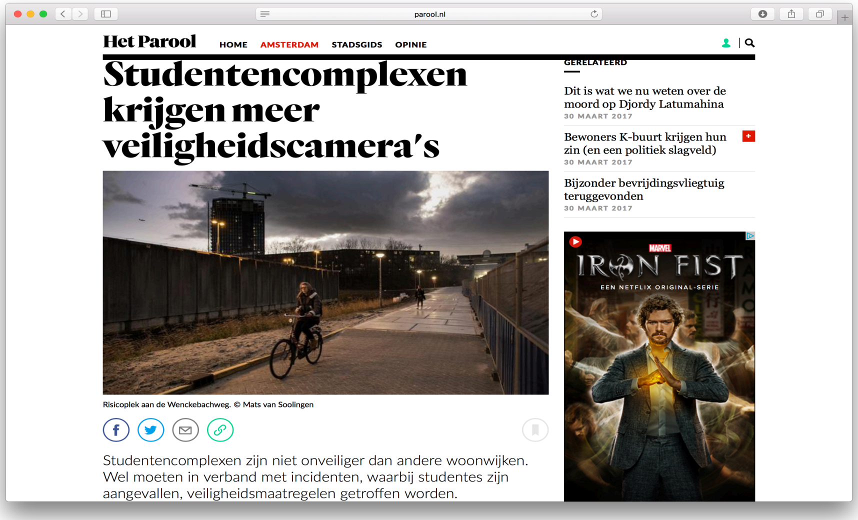
Video events will be used in many more new ways.

New surveillance

reconstruction



prevention


Het Parool HOME AMSTERDAM STADSGIDS OPINIE

Studentencomplexen krijgen meer veiligheidscamera's

Risicoplek aan de Wenckebachweg. © Mats van Soolingen

Studentencomplexen zijn niet onveiliger dan andere woonwijken. Wel moeten in verband met incidenten, waarbij studentes zijn aangevallen, veiligheidsmaatregelen getroffen worden.

GERELATEERD

Dit is wat we nu weten over de moord op Djordy Latumahina
30 MAART 2017

Bewoners K-buurt krijgen hun zin (en een politiek slagveld)
30 MAART 2017

Bijzonder bevrijdingsvliegtuig teruggevonden
30 MAART 2017

... and reused in old applications ...

New interaction

... and used as part of a loop.

Chen MM 2004

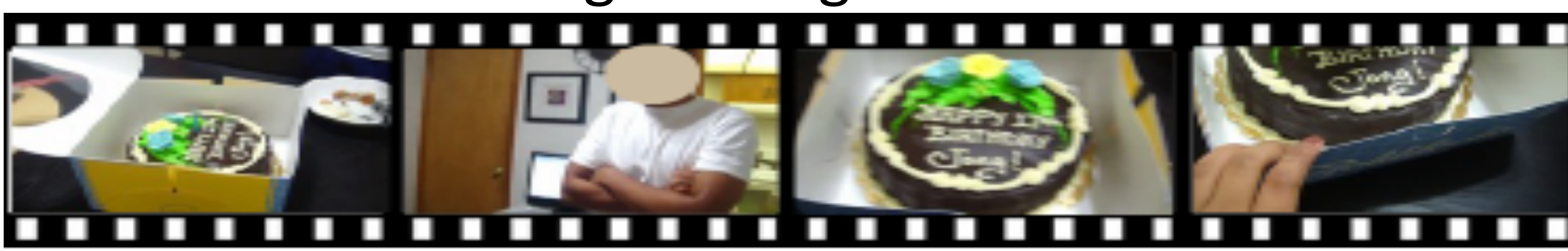
FOCUS

Video is filled with what?

Acts > Actions > Events - …

Acts: driving a screw

Actions: grooming an animal



Events: birthday party

Act > Action > Event > …

more degrees of freedom, **rigid schemes can't work**

act atomic motion pattern	*sleeping*
	running
	driving a screw
action functional pattern	*shaking hands*
	removing a lit
	serving the ball
event purposeful pattern	*serving an ace*
of actors, objects	*welcoming a friend*
and motions	*repairing an appliance*

Act > Action > Event > …

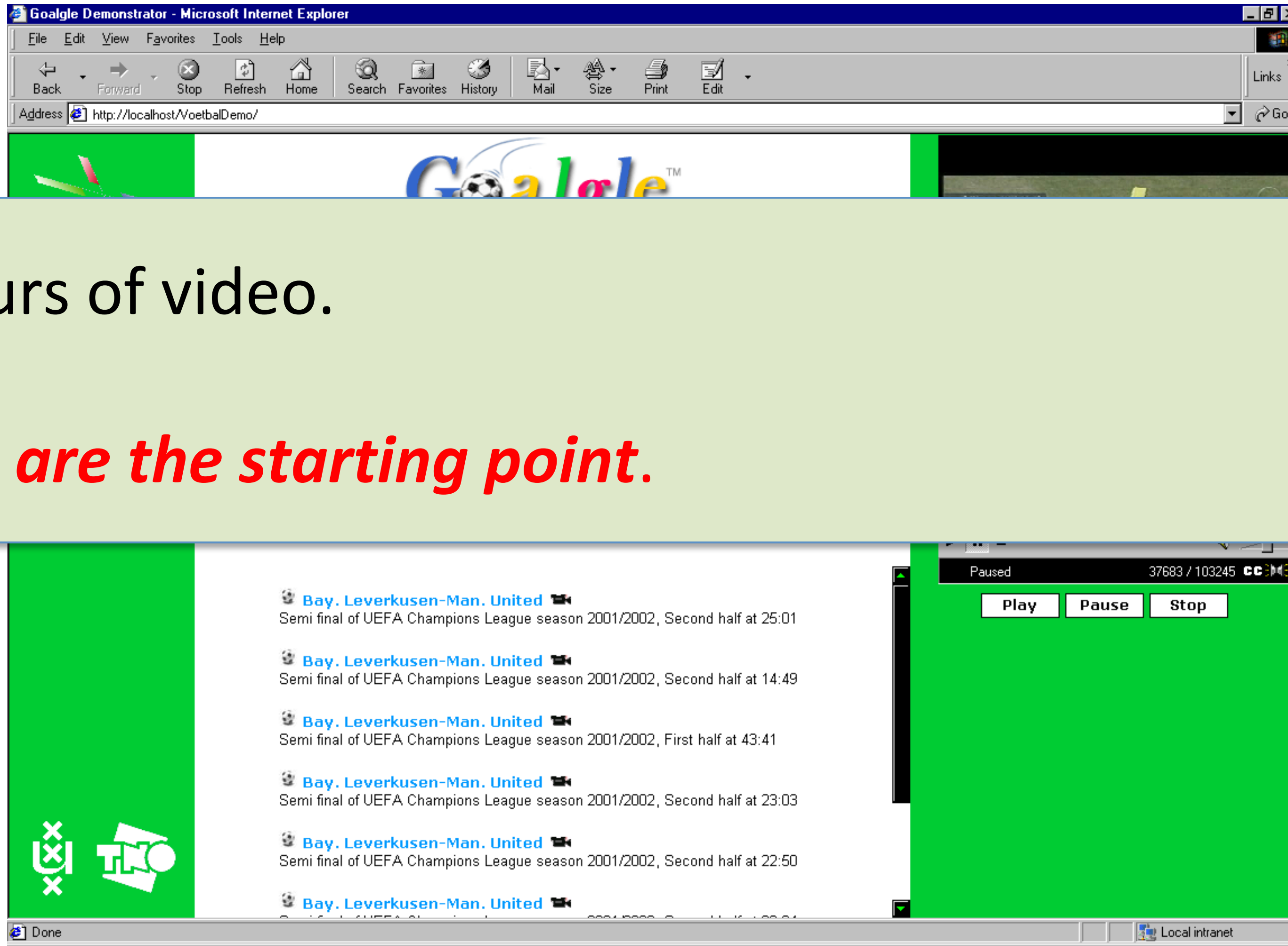
more degrees of freedom,
loss of order, of stage, of acts

		time frame	*pattern variations*
act	*sleeping*	±1s	pose
	running	±2s	dress, gait
	driving screw	±2s	repetition pace
action	*shaking hands*	2-5s	routine, active, soli
	removing a lit	2-5s	size, temperature
	serving ball	2-5s	camera
event	*serving an ace*	5-10s	camera, in
	welcoming	1-5m	choice acts, actions
	repairing	1-60m	choice acts, actions

DRIVING FACTORS

What makes success?

Goalgle



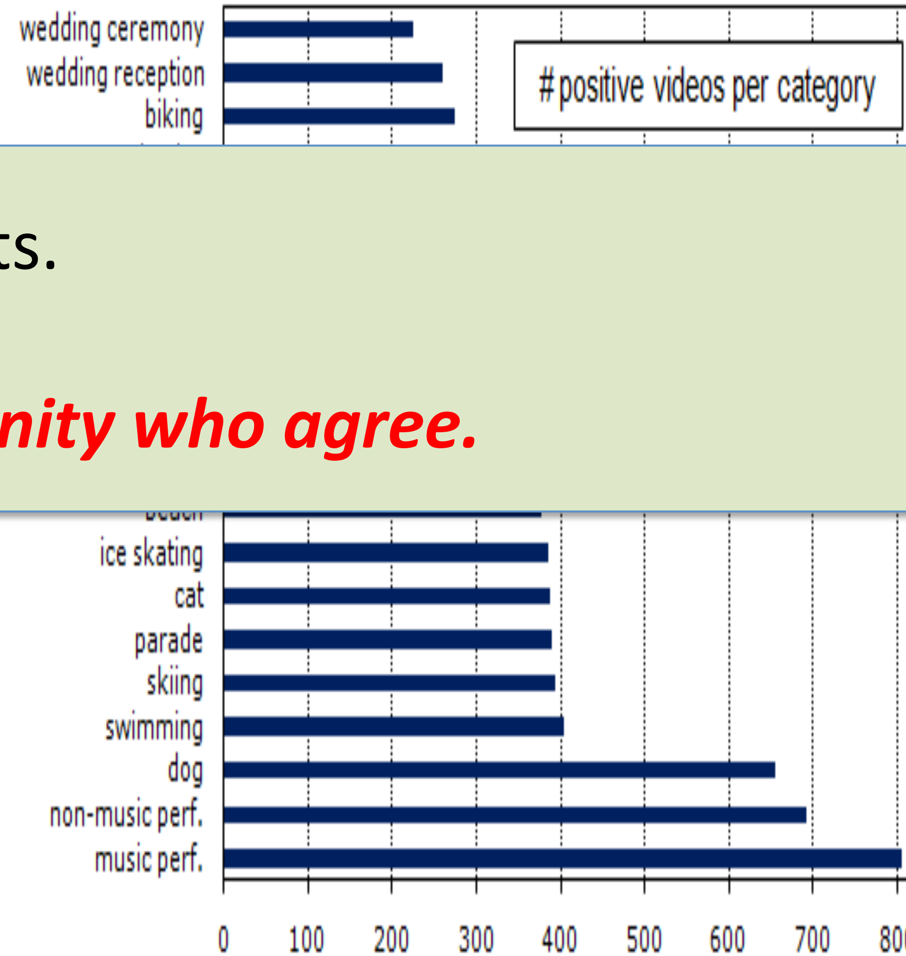
9 hours of video.

***Data are the starting point*.**

C.Snoek ICME 2003

CCV Columbia

- # videos: 9,317
  - (210 hrs in total)



Everyone their own results.

Progress needs a community who agree.

- average length
  - 80 seconds
- # defined categories
  - 20
- annotation method
  - Amazon Mechanical Turk

http://www.ee.columbia.edu/ln/dvmm/CCV/

TRECVID Internet video collections

Collection Name	Designated Uses	Target sizes	Annotation
Pilot	2010 Development collection	1,723 clips	Clip content annotation for both

The important moment

The driving factor is a shared & open competition.

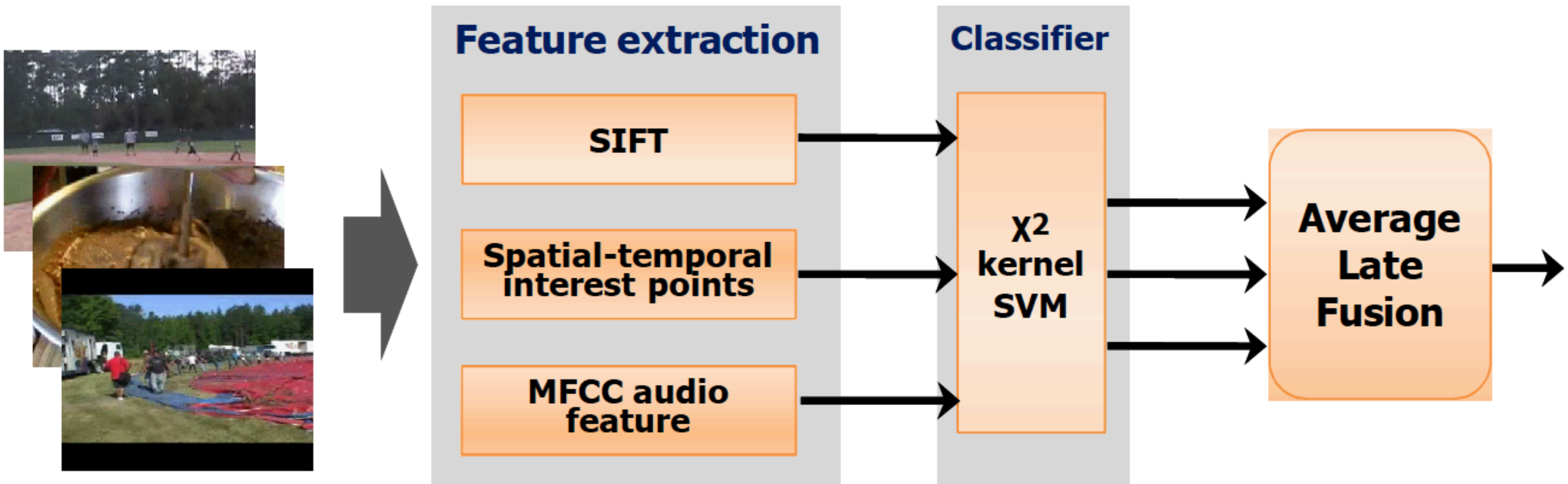
	2012-2015 (1) and (2) merged to a single training collection		Clip content annotation for the opaque subset
Progress	2012-2015: test collection	120K clips, 4000 hrs	No clip content annotation
Novel 1	2014: test collection	120K clips, 4000 hrs.	No clip content annotation
Novel 2	2015: test collection	120K clips, 4000 hrs.	No clip content annotation

CLASSIFICATION

Giving events a name, step by step TRECvid.

2010 Media diversity

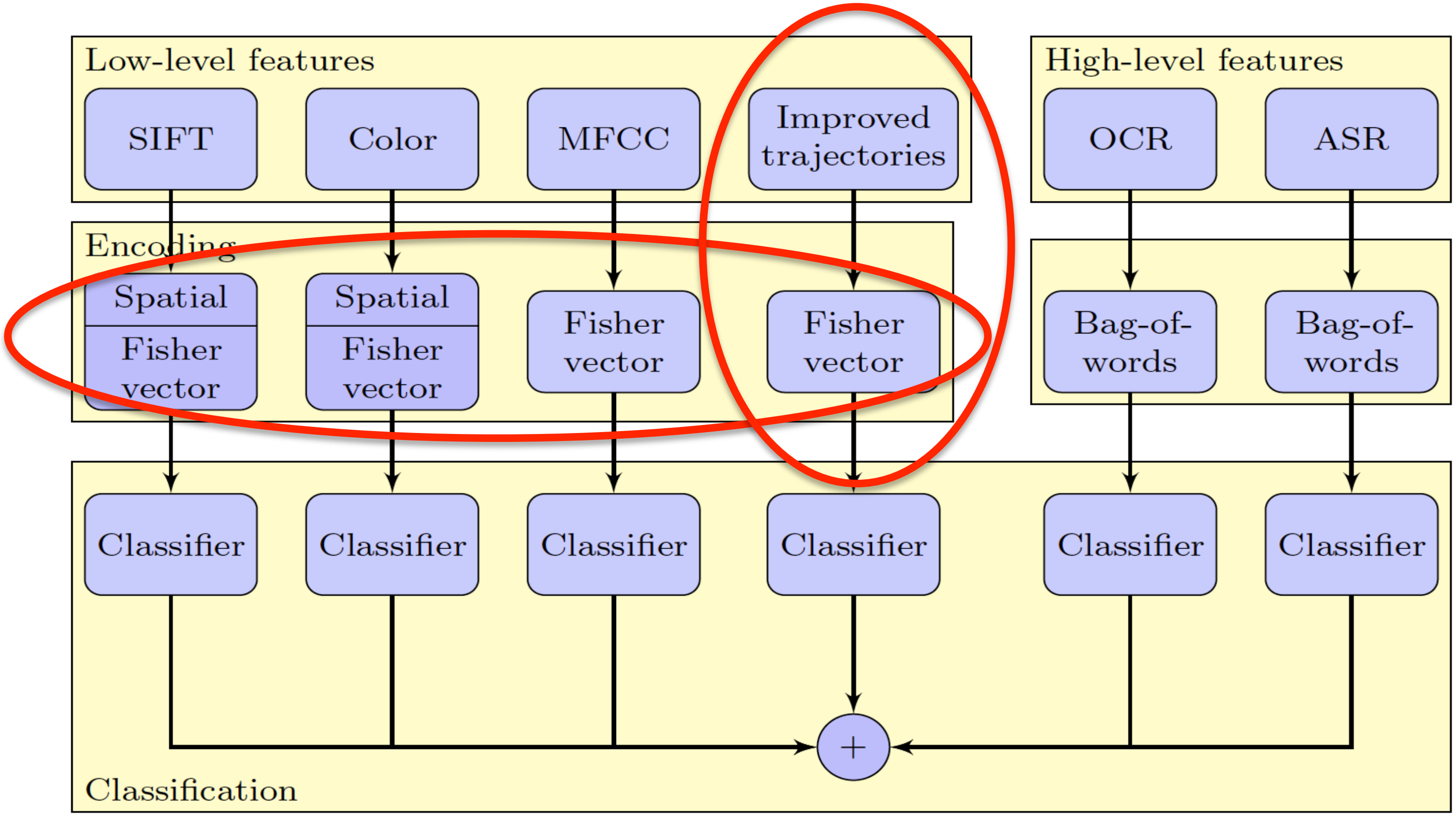
Diverse is better, more is better, fusion is better.



Y.G.Jiang TRECVID10 P.Natarjan CVPR12 Wang ICCV13 *others*

Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Subh Bhattacharya, Dan Ellis, Mubarak Shah, Shih-Fu Chang, **Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching**, NIST TRECVID Workshop, 2010.

2012/13 Trajectories and aggregation



Dense trajectories are more and Fisher aggregation are fusion.

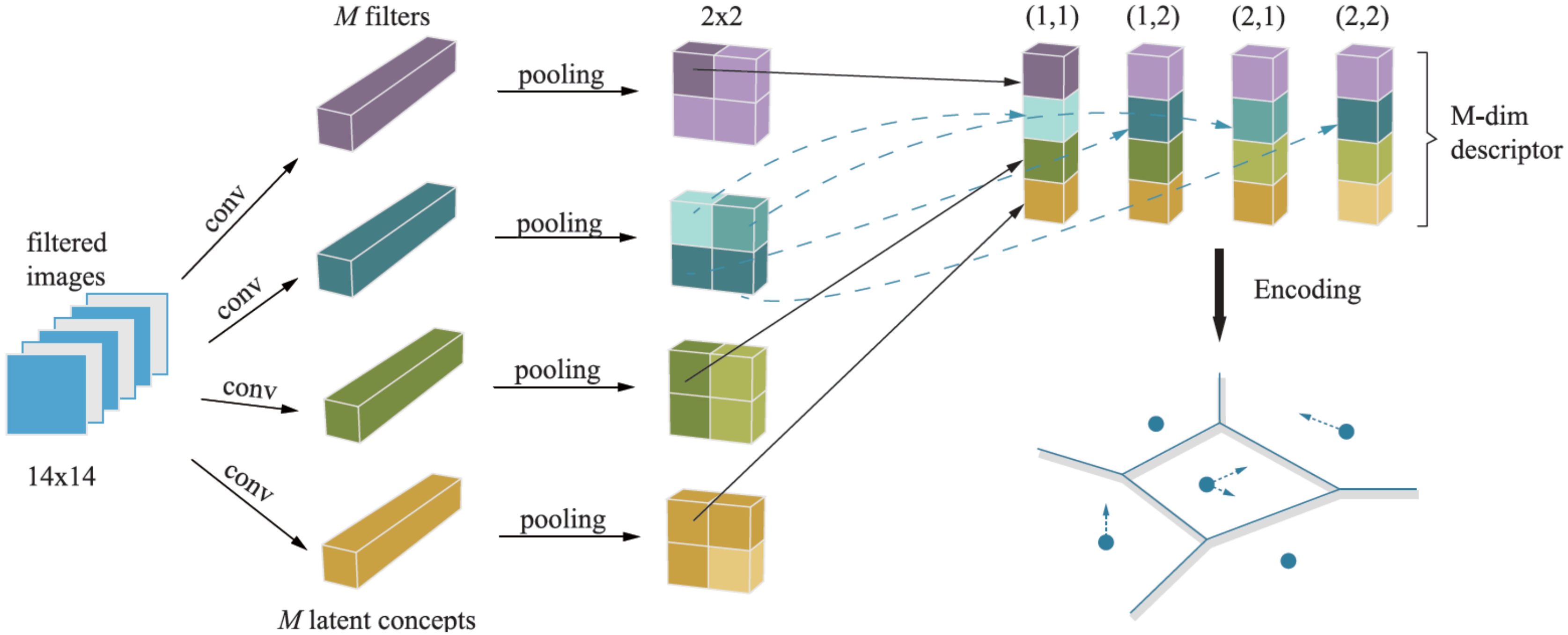
This is the end of hand engineered features.

INRIA LEAR H.Wang CVPR 2011

2014 Deep learning & VLAD

Networks integrate features and classifiers.

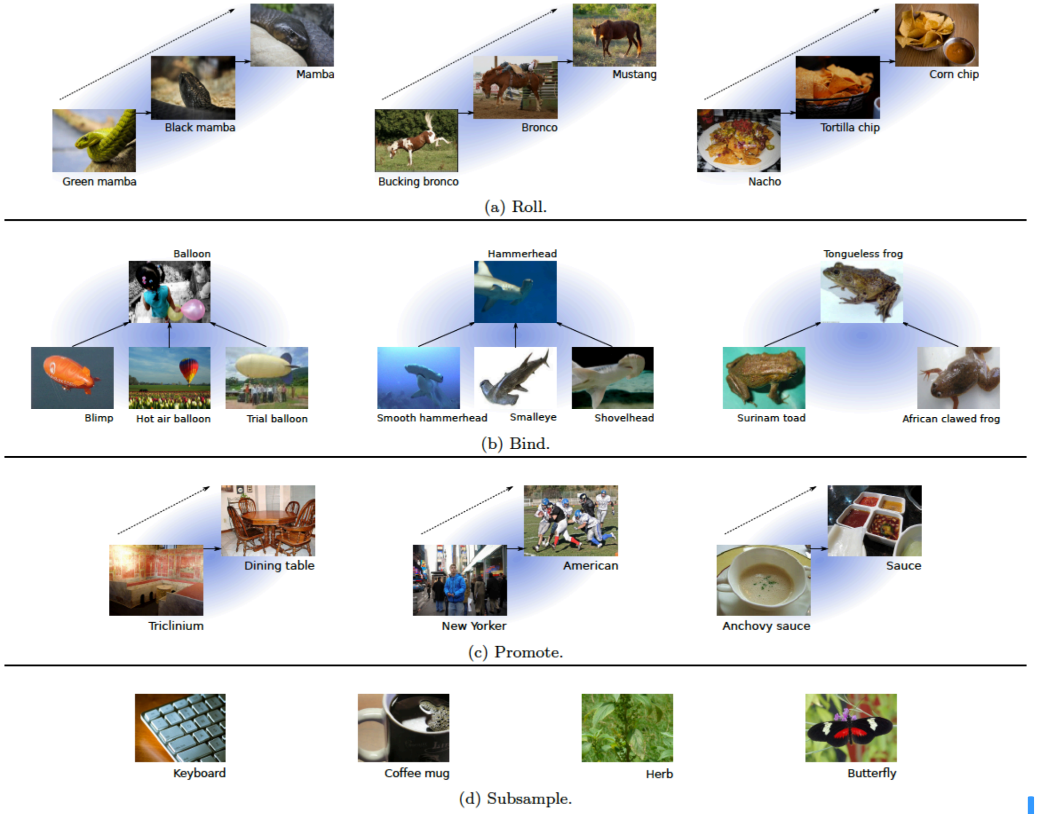
Deep learning builds in fusion of diverse, more and late.



CMU Xu CVPR 2015

2015 Prior knowledge

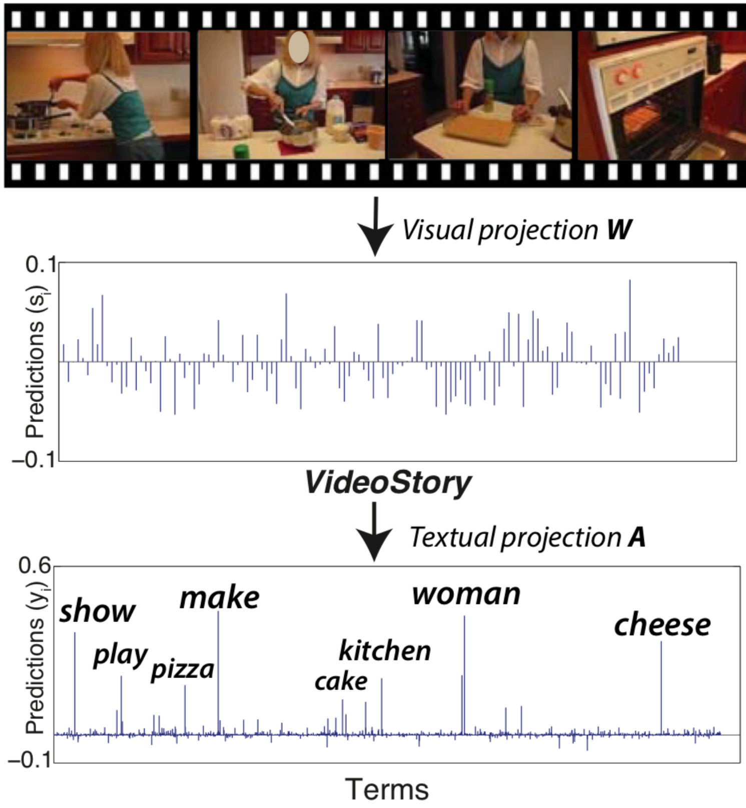
Insert 15000 ImageNet detectors pruned, but first reorganize prior knowledge removing fine semantics and merging small sets.



UvA P.Mettes ICMR 2016

2016 Joint embedding

Fuse media diverse in one embedding to compose stories



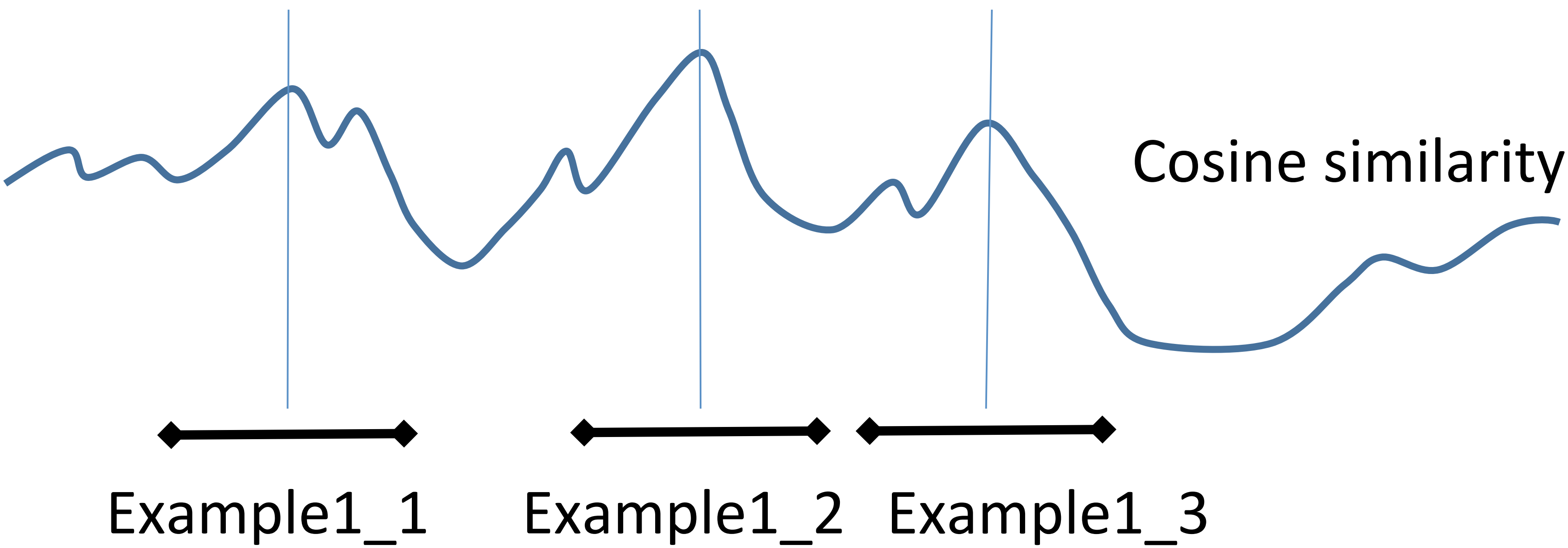
Pre-train representation on webly-supervised videos

.. detectors selected for generality and specificity.

… to achieve stories, even when that class has few data.

UvA A.Habibian MM 2014

2017 Expand training material



More by less data

UvA D. Koelma 2017

RETRIEVAL

Every question is new, so classification not for every day.

Event retrieval: is it zero-shot?

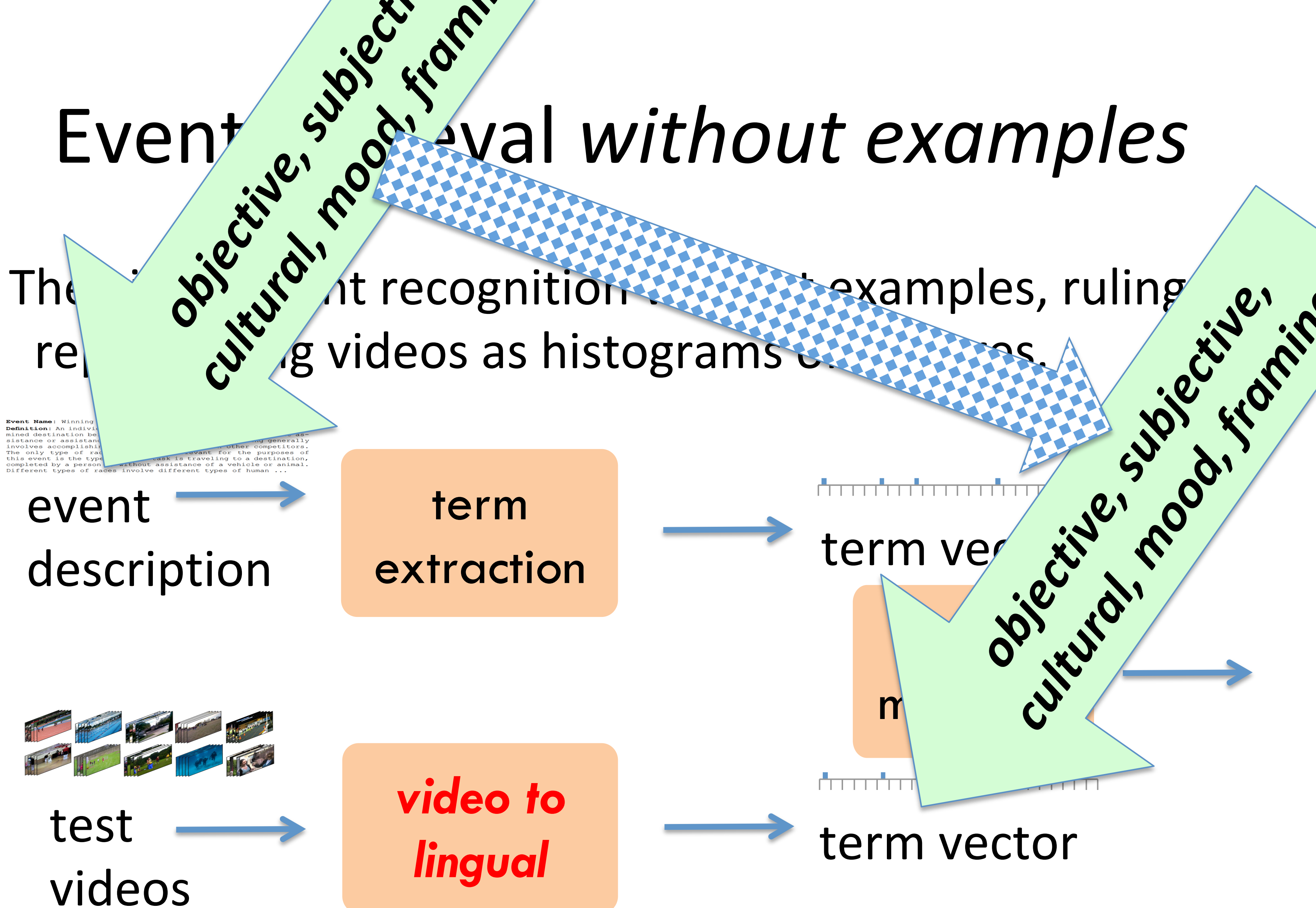
Zero shot aims to classify test videos by predefined mutual relationship using class-to-attribute mappings

We aim for a new event by a text only.

Lampert PAMI 2013 *others*

Event ... eval *without examples*

The ... nt recognition ... examples, ruling re... g videos as histograms o... es.



Jiang TRECVID 2010 Natarajan CVPR 2012 Wang ICCV 2013

Nouns are easy, propositions are not

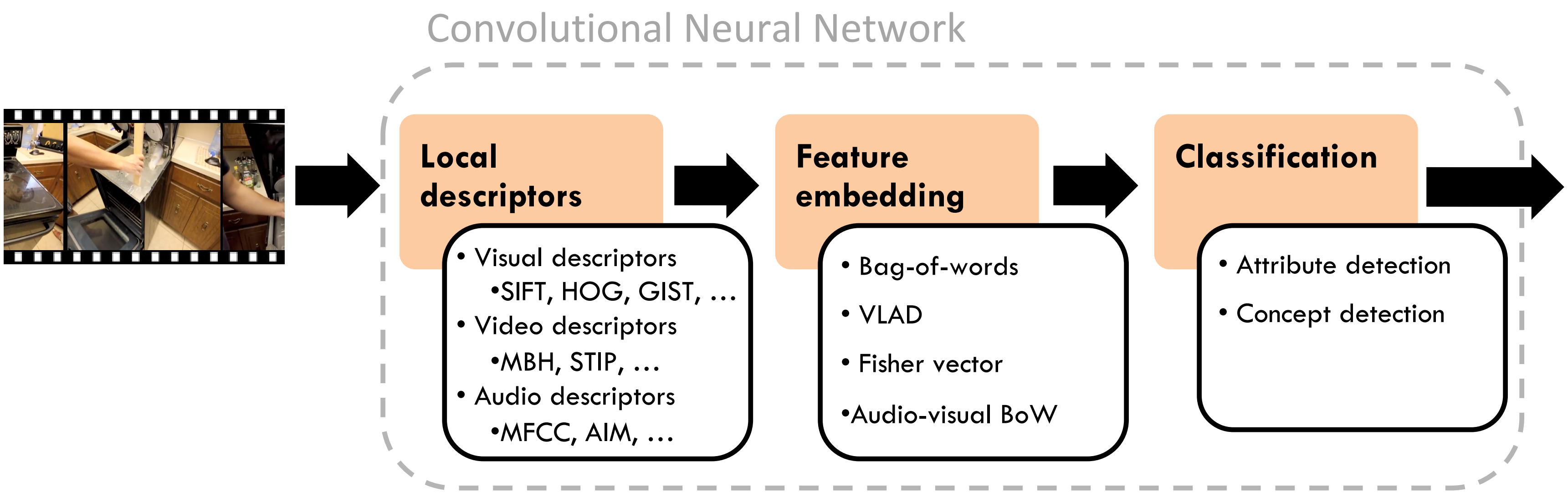
Can we have a vote?

Nouns are stable, adjectives personal

Old is visually different for every notion.

Concept embedding for retrieval without examples

Representing videos as histograms of concept scores



Problem: one classifier against the complexity of the world.

Wu CVPR 2014 A.Habibian ICMR 2014

CONCEPT EMBEDDING

Concept embedding label expansion

Expanding the labels by logical combinatorics,

Label expansion expands the vocabulary for free:

bike .and. road for *bicycle trick*, not bike .or. road.

Habibian ICMR 2014

Concept embedding qualitative results

Top ranked videos for *flash mob gathering.*

Most important concepts in their video representation

Detected Videos	Composite Concepts
	Group-AND-Dance-AND-Shopping Celebrating-OR-Marching Performance-OR-Music People-OR-Girl Surprise-OR-Party
	Group-AND-Dance-AND-Shopping Band-OR-Singining Inside-OR-School Performance-OR-Music Surprise-OR-Party
	Group-AND-Dance -AND-Shopping Practice-OR-Gym Living-AND-Room Street-OR-Inside Performance-OR-Music

Still need a labeled basis for each concept classifier.

Habibian ICMR 2014

VIDEO TO TAG-TERM EMBEDDING

Embedding inspiration from tags

Embedding based on free social tagged videos only, without the need for training any intermediate detectors.

Inspired by:

Xirong Li TMM 2009

Video2vec embedding

Can we learn the embedding from videos and their stories?

Video

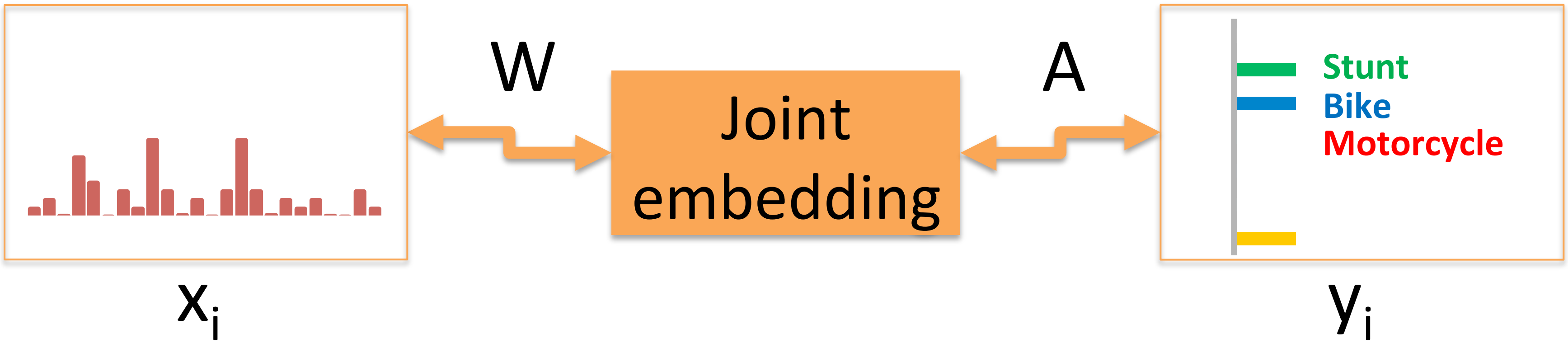
Story

Story usually highlights the key concepts in video jointly.

Videos and stories are freely available on YouTube.

A.Habibian ACM MM 2014 A.Habibian PAMI 2017

Video2vec embedding



Joint space where $x_i W \approx y_i A$.

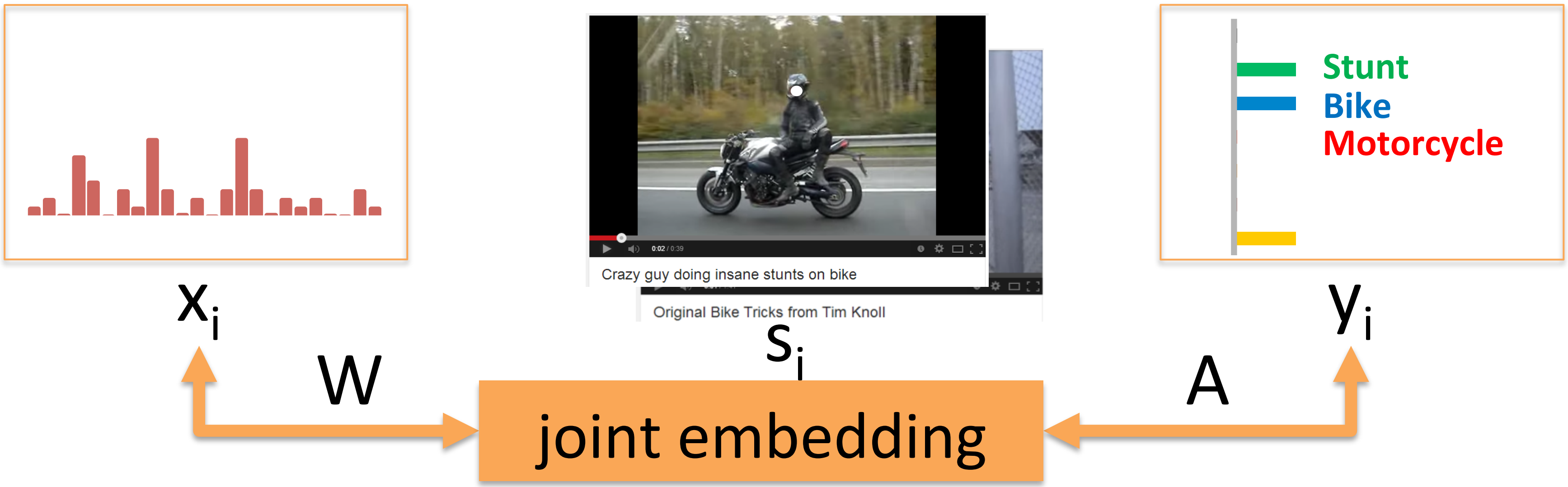
Explicitly relate training W and A from multimedia.

W = Visual projection matrix   individual term classifiers

A = Textual projection matrix   select/group terms

Rasiwasa MM 2010 Weston IJCAI 2011 Akata CVPR 2013

Video2vec embed the video story



Learn W and A such that *descriptiveness* preserves video descriptions and *predictability* recognizes terms from video content

A.Habibian ACM MM 2014 A.Habibian PAMI 2017

Video2vec key observation

By grouping terms, the number of classes is reduced.
For training classifiers, more positives needed per group.
We can train from freely available web data.

Video2vec joint optimization

S is (the size of) the embedding

L_d Loss function for descriptiveness.

L_p Loss function for predictability.

$$L_{\mathrm{VS}}(\boldsymbol{A}, \boldsymbol{W}) = \min_{\boldsymbol{S}} L_d(\boldsymbol{A}, \boldsymbol{S}) + L_p(\boldsymbol{S}, \boldsymbol{W})$$

Jointly optimize descriptiveness and predictability.

A.Habibian ACM MM 2014 A.Habibian PAMI 2017

Video2vec objective descriptiveness

The Video2vec embedding should be descriptive.

$$L_d(\boldsymbol{A}, \boldsymbol{S}) = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{y}_i - \boldsymbol{A}\boldsymbol{s}_i\|_2^2 + \underbrace{\lambda_a \Omega(\boldsymbol{A}) + \lambda_s \Psi(\boldsymbol{S})}_{\text{Regularizers}}$$

Original transcriptions

Regularizers

Reconstructed terms

Latent semantic indexing with L2 norm.

A.Habibian ACM MM 2014 A.Habibian PAMI 2017

Video2vec objective predictability

The Video2vec embedding should be predictable.

$$L_p(\boldsymbol{S}, \boldsymbol{W}) = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{s}_i - \boldsymbol{W}^\top \boldsymbol{x}_i\|_2^2 + \lambda_w \Theta(\boldsymbol{W})$$

Video2vec embedding

Regularizer

Video feature embedding

A.Habibian ACM MM 2014 A.Habibian PAMI 2017

Video2vec 46K dataset

Videos and title descriptions from higher quality YouTube,

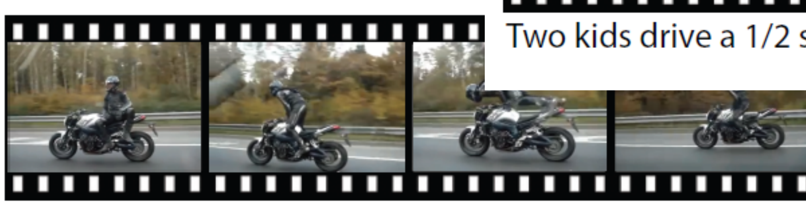
46K videos, 19K terms in description.

Features x_i any combination.

Seeded from video event descriptions y_i in bags.

Cute tabby cat gives her dog a bath

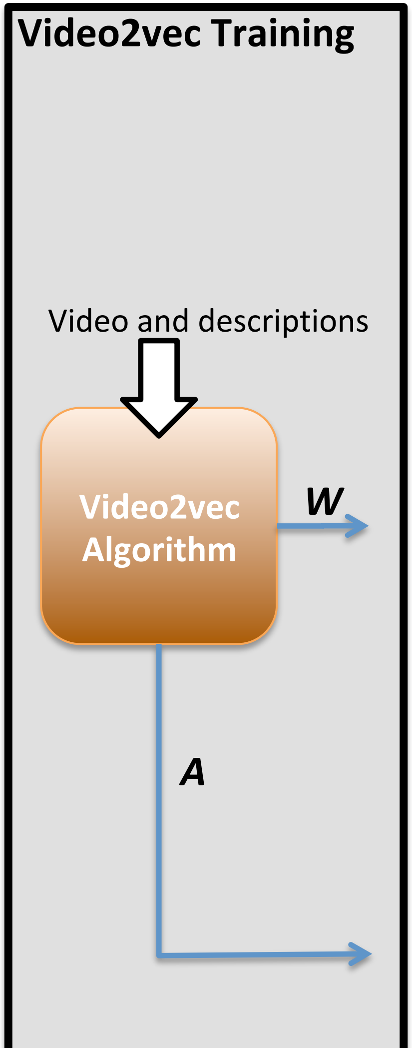
Two kids drive a 1/2 size Jeep through mud



Crazy guy doing insane stunts on bike.

Available for download: www.mediamill.nl

Video2vec training method



Stochastic Gradient Descent starting from a random sample.

The sample gradient wrt objective is:

$$\nabla_{\boldsymbol{A}} L_{\mathrm{VS}} = -2\,(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{s}_t)\,\boldsymbol{s}_t^{\top} + \lambda_a \boldsymbol{A},$$

$$\nabla_{\boldsymbol{W}} L_{\mathrm{VS}} = -2\,\boldsymbol{x}_t \left(\boldsymbol{s}_t - \boldsymbol{W}^{\top}\boldsymbol{x}_t\right)^{\top} + \lambda_w \boldsymbol{W}, \text{ and}$$

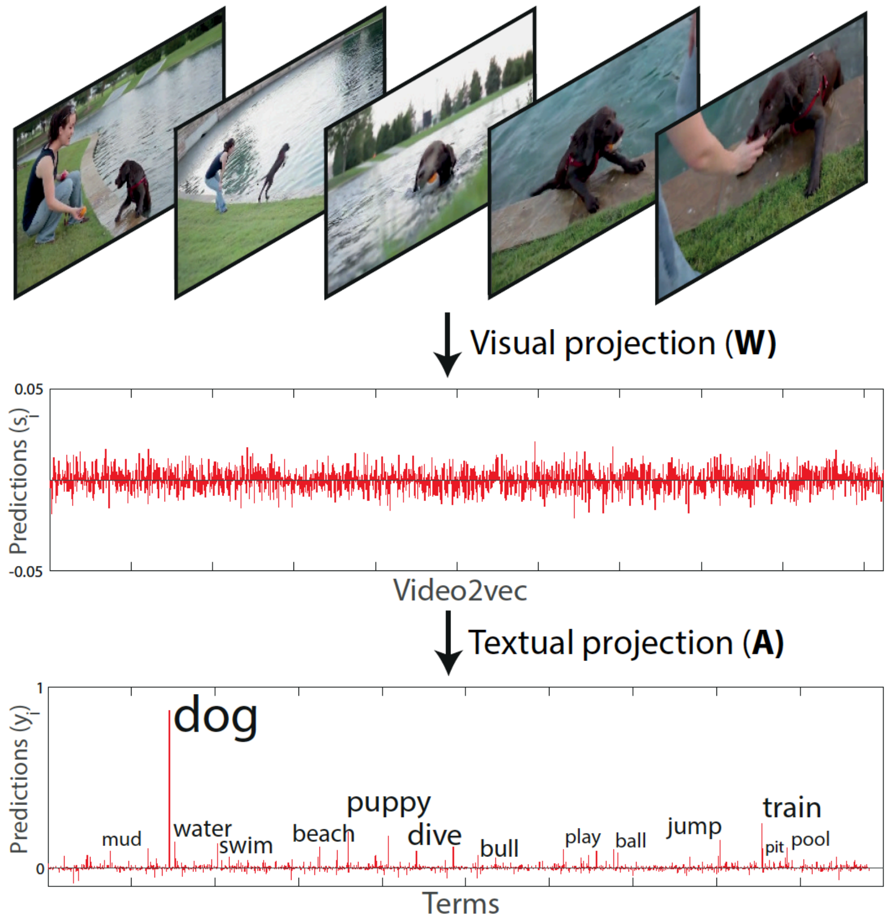
$$\nabla_{\boldsymbol{s}_t} L_{\mathrm{VS}} = \;\; 2\left[\boldsymbol{s}_t - \boldsymbol{W}^{\top}\boldsymbol{x}_t - \boldsymbol{A}^{\top}(\boldsymbol{y}_t - \boldsymbol{A}\boldsymbol{s}_t)\right] + \lambda_s \boldsymbol{s}_t.$$

Update parameters with step-size η.

Start **A** and **S** from SVD of term vectors **Y**.

Bottou ICCS 2010

Video2vec at work



1. Project visual features

$$\boldsymbol{s}_i = \boldsymbol{W}^\top \boldsymbol{x}_i,$$

2. Translate to text

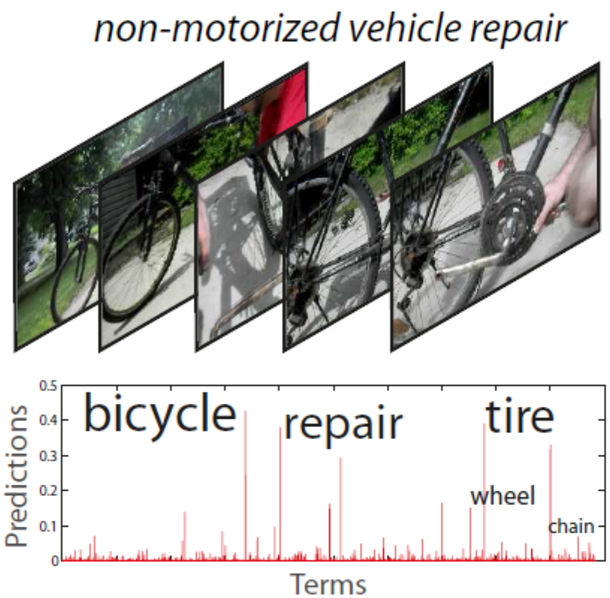
$$\hat{\boldsymbol{y}}_i = \boldsymbol{A}\boldsymbol{s}_i,$$

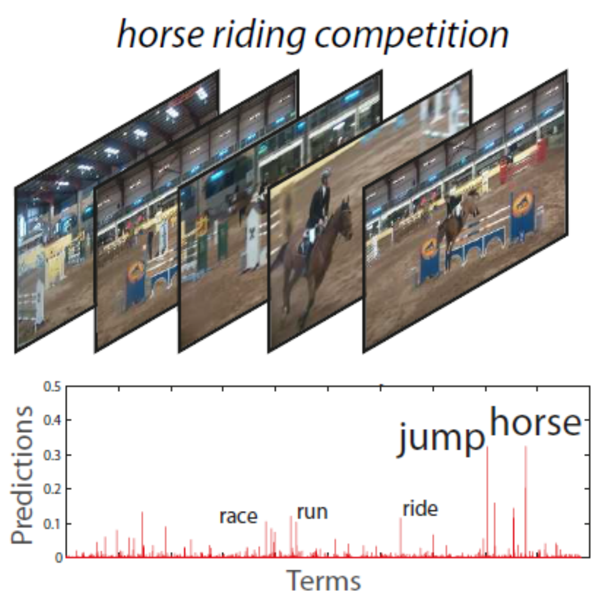
3. Cosine distance match

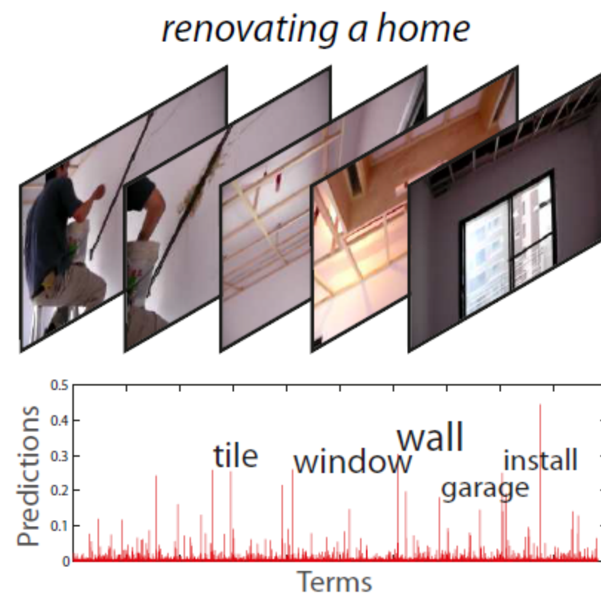
$$s_e(\boldsymbol{x}_i) = \frac{\boldsymbol{y}^{e\top} \hat{\boldsymbol{y}}_i^e}{||\boldsymbol{y}^e|| \quad ||\hat{\boldsymbol{y}}_i^e||}$$

A.Habibian ACM MM 2014 A.Habibian PAMI 2017

Video2vec predicted terms



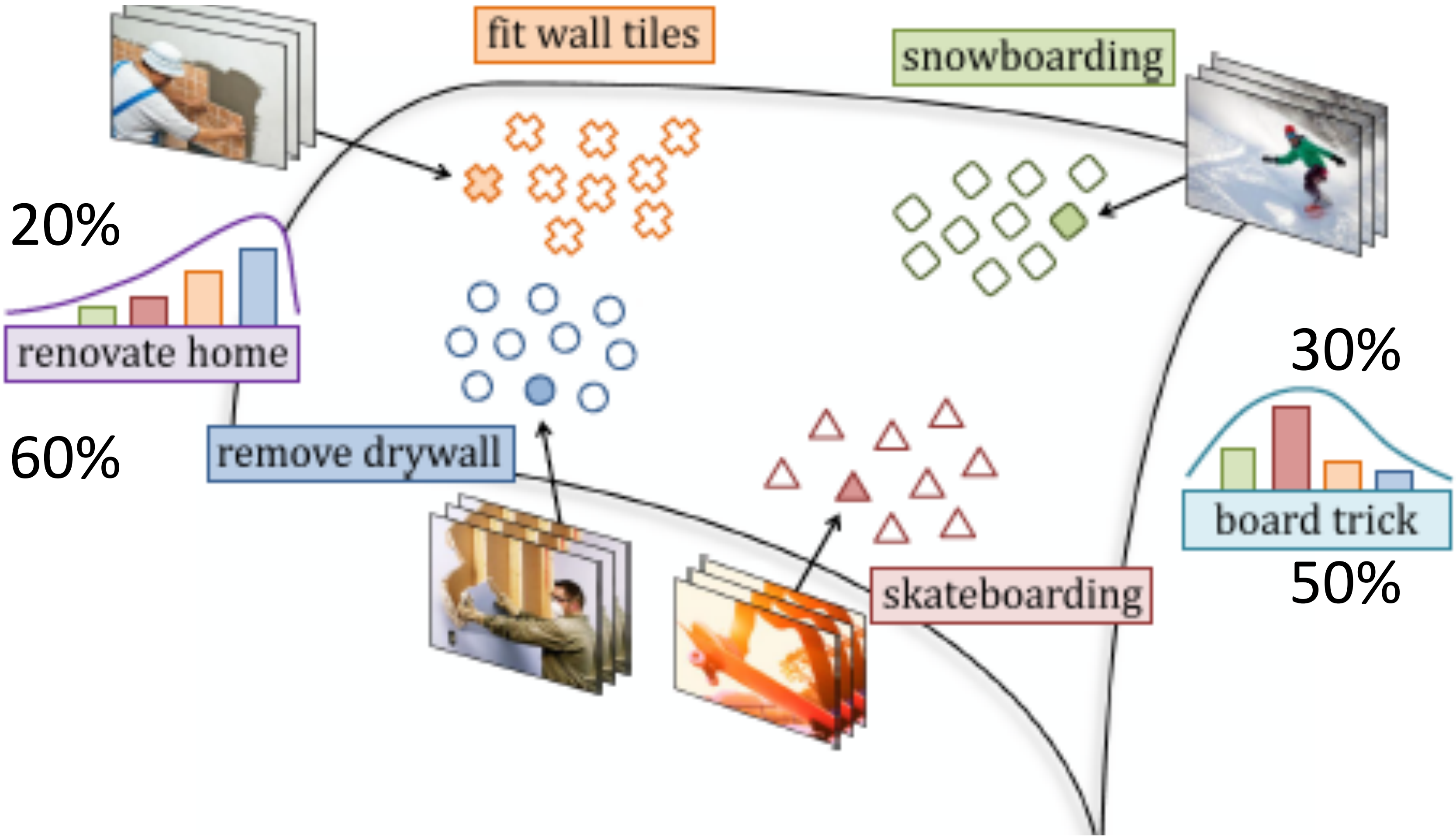




A.Habibian ACM MM 2014 A.Habibian PAMI 2017

UNIFIED METRIC EMBEDDING

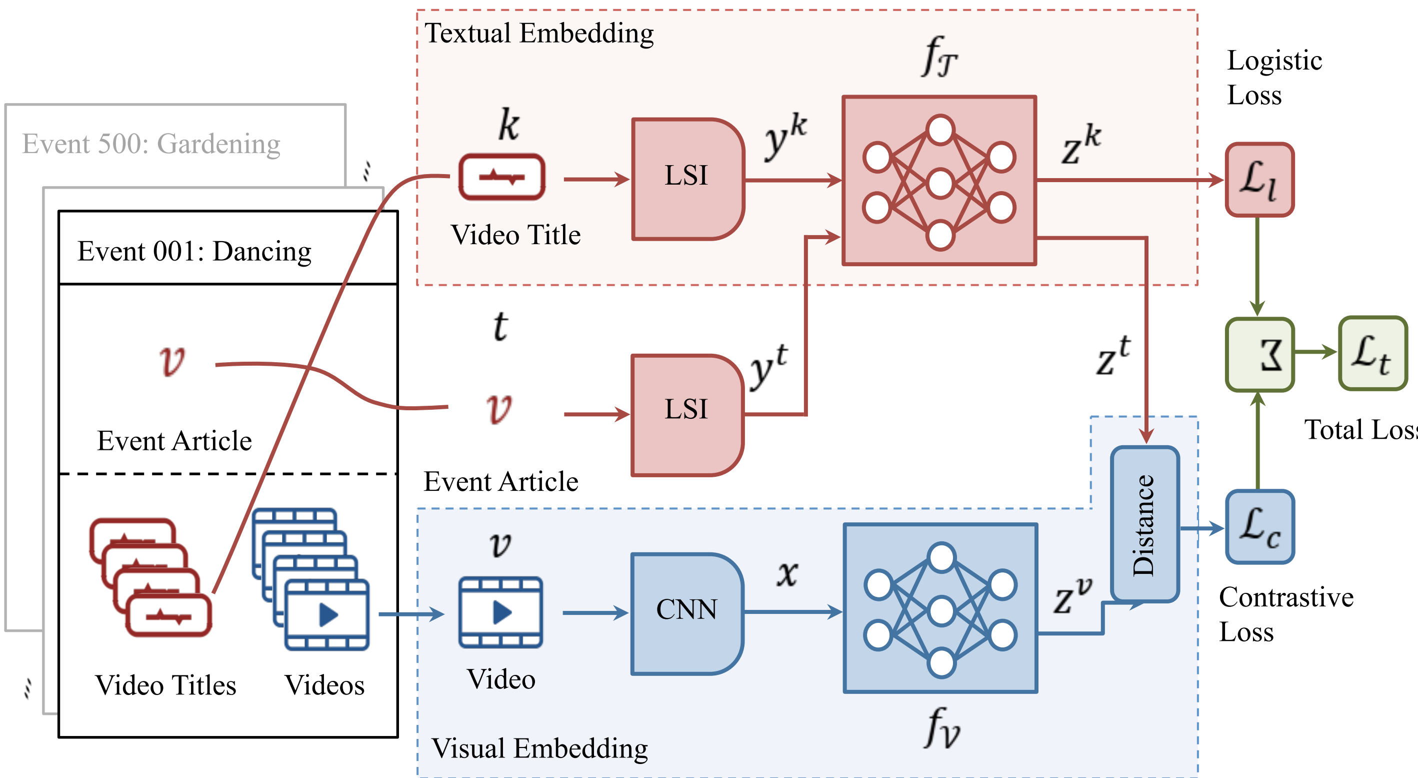
Unified metric embedding



N.Hussein CVPR2017

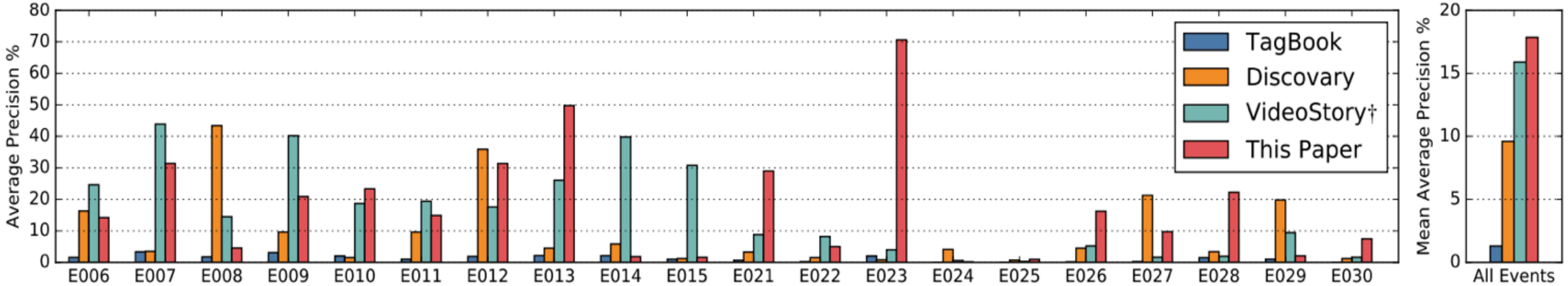
Unified metric embedding

Zero-exemplar is learning from pre-defined events plus novel ones as a probability over the existing events.

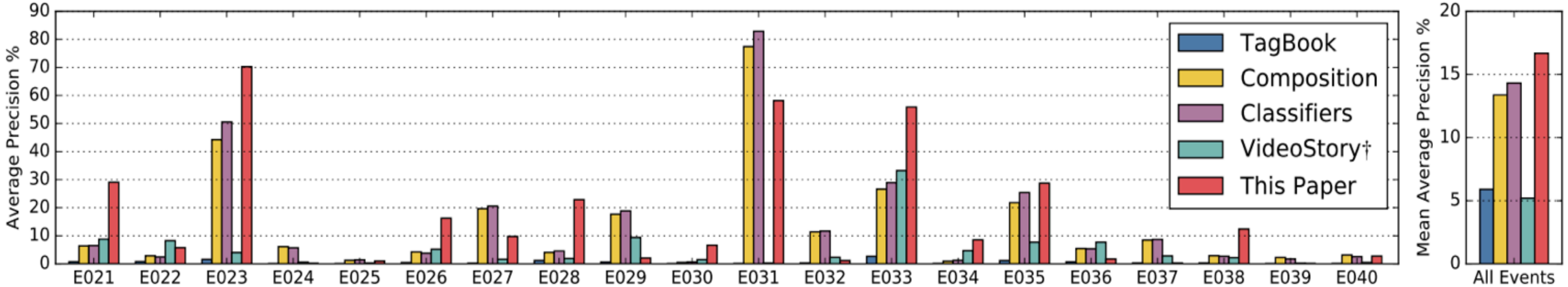


N.Hussein CVPR2017

Unified metric embedding quantitative



(a) MED-13 Dataset



(b) MED-14 Dataset

N.Hussein CVPR2017

Unified metric embedding qualitative

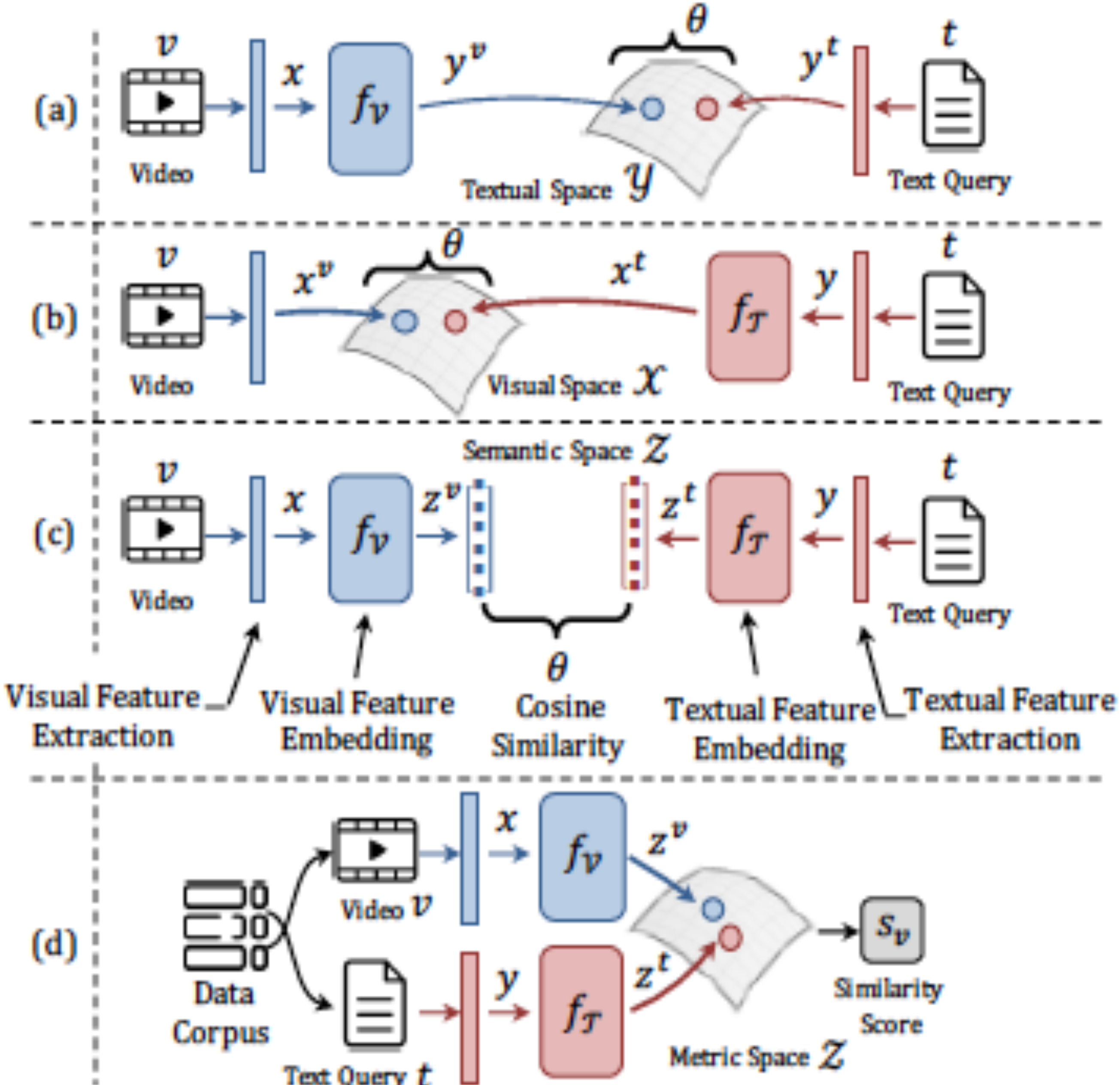
Renovating home improve a home by rebuilding parts of the structures.

Success

Failure

N.Hussein CVPR2017

EVENT RECOGNITION BY EMBEDDING



N.Hussein CVPR2017

Retrieval by embedding results

Authors		Published	mAP
Habibian et al.	concept embedding	ICMR 2014	6.4
Ye et al.		MM 2015	9.0
Mazloom et al.		ICMR 2015	11.9
Wu et al.		CVPR 2014	12.7
Jiang et al.		AAAI 2015	12.9
Mazloom et al.	tag embedding	TMM 2016	12.9
Liang et al.	big data & reranking	MM 2015	18.3
Habibian et al.	joint embedding	TPAMI 2017	20.0
Hussein et al.	unified metric embed	CVPR 2017	17.9

N.Hussein CVPR2017 A.Habibian PAMI 2017

OTHER CHALLENGES

In the kitchen of the future.

TRECVID SURVEILLANCE EVENTS > TRECVID ACTIVITIES EXTENDED VIDEO

slides by Jon Fiscus (NIST)

COMPLEX ACTIVITIES IN VIDEO

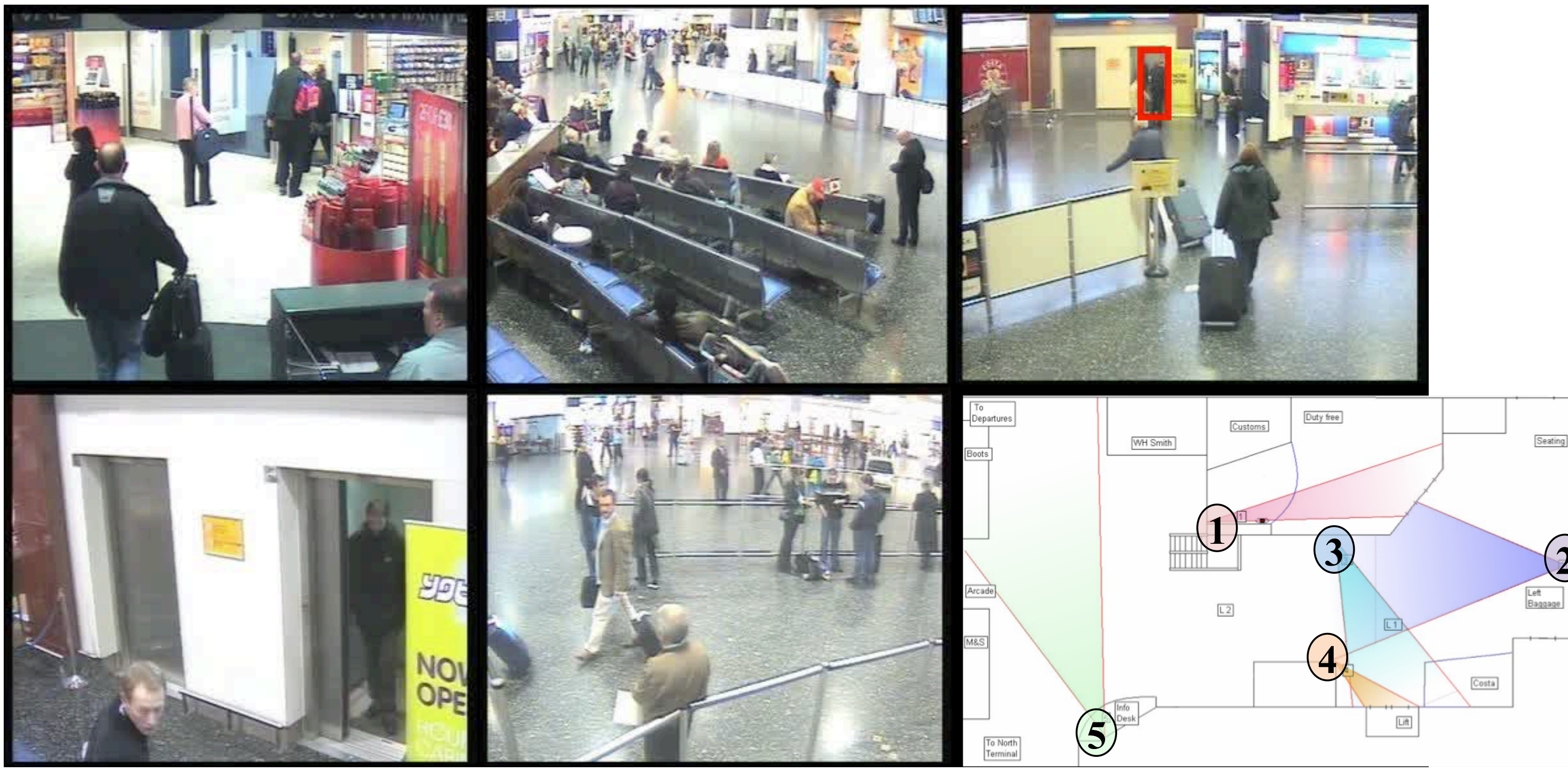
at the UvA by N.Hussein, S. Gavves, C. Snoek *others*

Multi-cam surveillance from text

Controlled Access Door 1

2 Waiting Area

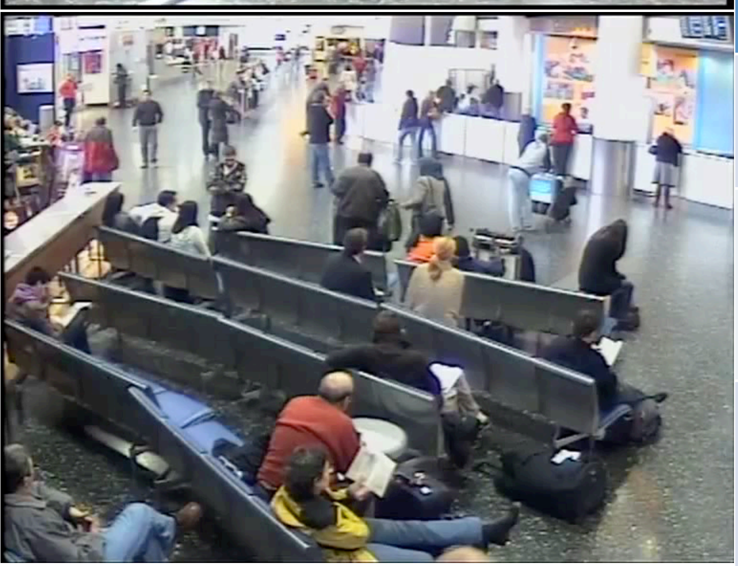
3 Debarkation Area



4 Elevator Close-Up

5 Transit Area

Events of Interest

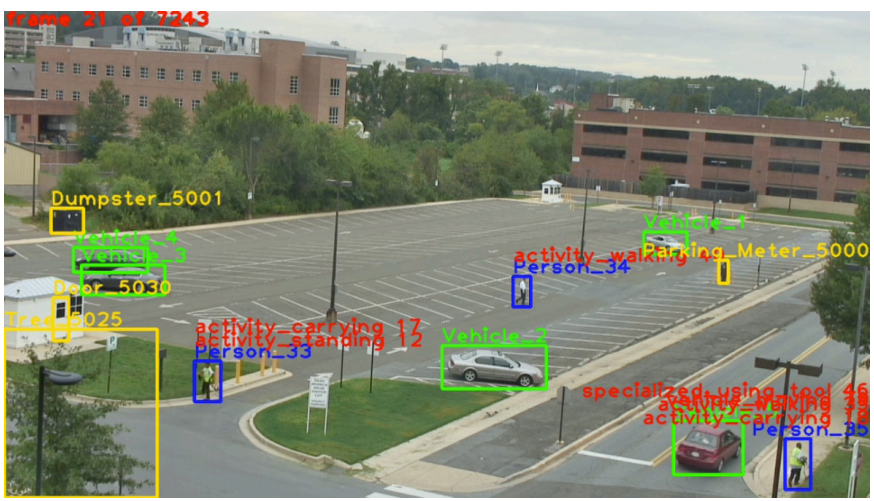


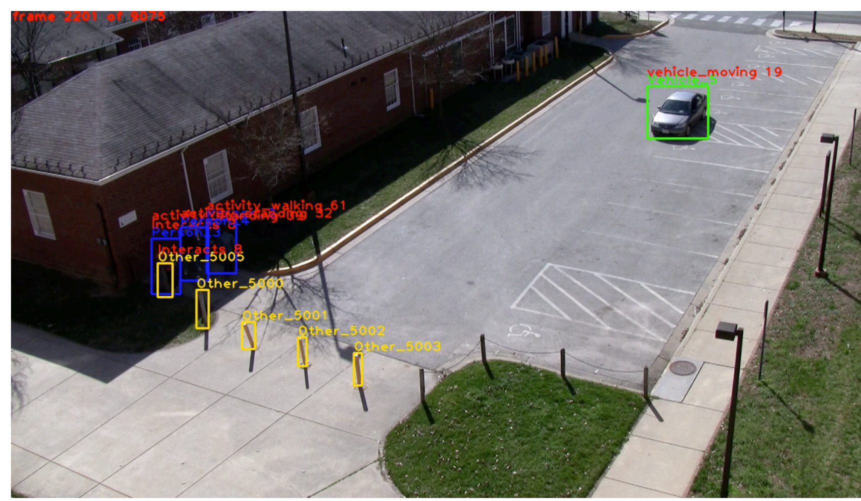
Single Person events	
PersonRuns	Someone runs
Pointing	Someone points
Single Person + Object events	
CellToEar	Someone puts a cell phone to his/her head or ear
ObjectPut	Someone drops or puts down an object
Multiple People events	
Embrace	Someone puts one or both arms at least part way around another person
PeopleMeet	One or more people walk up to one or more other people, stop, and some communication occurs
PeopleSplitUp	From two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the

ActEV new task per 2018

Successor of Surveillance Event Detection by adding a large collection of multi-camera video data, both of simple and complex activities.

ActEV will address activity detection for both forensic applications and for real-time alerting.





Recognizing complex tasks

Strong temporal models are no longer valid.

This depicts *cooking food* regardless frame order.

Epic kitchen